DOCUMENT RESUME

ED 037 358 SE 008 113

AUTHOR Benson, Bernard W.
TITLE The Development and Implementation of an Instrument to Assess Cognitive Performance in High School Biology.
INSTITUTION Tennessee Univ., Chattanooga.
PUB DATE Mar 70
NOTE 41p.; Paper presented at Annual Meeting of the National Association for Research in Science Teaching (43rd, Minneapolis, Minne., March 5-8, 1970)

EDRS PRICE EDRS Price MF-$0.25 HC-$2.15
DESCRIPTORS *Biology, *Cognitive Tests, *Critical Thinking, *Evaluation, *Secondary School Science, Test Construction, Test Reliability

ABSTRACT

An instrument, "The Assessment of Cognitive Transfer in Science Inventory," was designed to evaluate cognitive performance in biology. The instrument is based on students' verbal responses to a structured sequence of situations and questions. Items were classified in terms of a modification of Bloom's Taxonomy of Educational Objectives. The instrument was tested on students sampled to represent a cross-section of biology instruction. Analysis of the data includes tests of independence of the classification categories. Estimations are given of item difficulty, discrimination, and test reliability. (EB)

THE DEVELOPMENT AND IMPLEMENTATION OF AN INSTRUMENT TO ASSESS COGNITIVE PERFORMANCE IN HIGH SCHOOL BIOLOGY

by

Bernard W. Benson

The University of Tennessee at Chattanooga

A Paper Presented at the Forty-third Annual Meeting of the National Association for Research in Science Teaching

March 6, 1970

Minneapolis, Minnesota

The Development and Implementation of an Instrument to Assess Cognitive Performance in High School Biology

As means or as ends the goals of secondary science instruction suggest increased emphasis on developing cognitive skills, as defined by Bloom (2), above the knowledge level. The proponents of said goals would agree with Kochendorfer's (12) premise that the ultimate test of any new "curriculum" is the extent to which it meets its desired goals.

Research at the operational level that demonstrates the effectiveness of such programs and identifies the factors associated with goal attainment, i.e., dependent variables, have taken a host of forms and reflect considerable heterogeneity in design and scope. Ramsey and Howe (14) discussed the problems encountered in such research. They provide a classification scheme that reflects the status of research in science education. Noticeably absent are entries that attempt to elucidate the transfer of training and learning. Sparcely represented in the research literature, are schemes which assess student achievement at the higher cognitive levels such as analysis, synthesis and evaluation, using evaluation instruments designed to measure student performance on tasks related in content, process, and strategy to the new courses.

Statement of the Problem and Purpose of the Study

Ennis (7) and others (13, 18) recognized the need for critical thinking tests in various subject matter areas. It is notorious in his words that "some are good critical thinkers in one area and not in other areas". Therefore, critical thinking to some extent (the extent not definable) is specific to the field in which it takes place. How then can one justify using measures like the Watson-Glaser Critical Thinking Appraisal (19) which do not refer to a specific science discipline or, for that matter, are not restricted to science? With such instruments serving as criterion measures we may well be assessing gains mediated by social studies or English classes or possibly television commercials.

The operational definition for critical thinking as used in this study is as follows: If we state explicit behaviors we expect of students as exemplars of the processes of science and at the same time state minimum levels of performance by cognitive level, students will be thinking critically if these behaviors are exhibited. In other words critical thinking cannot be separated from the act of cognition, if the act is one of analysis, synthesis, or evaluation. Those situations which capitalize on the structure and content of antecedent tasks or conceptualizations based on past experience will provide the best setting for defining or delineating such behaviors. An instrument designed to measure critical thinking

in a specific science area would be contingent on the continuation of the premise that each operation be based on the structure and content of antecedent tasks.

This investigator further submits that one cannot measure such entities as critical thinking or strategies of discovery within a discipline without designing instruments that measure other aspects of performance, specifically performance in thinking cognitively. For this reason the above definition of critical thinking, as applied herein, is contingent upon direct measures of cognition. The total learning process should be reflected in such proposed evaluation instruments.

If the Taxonomy of Educational Objectives: Cognitive Domain (2) is directly based on learning theory and the psychological processes involved in learning, it, or schemes developed from it, should prove valid tools for determining relationships among similar learning processes and between teaching methodology and learning process. As stressed by Bloom (3), future research which makes use of the taxonomy may reveal psychological relations among the different classes of objectives and the extent to which transfer and retention differ among the major types of objectives.

A corollary to the above definition might be that groups of operations be preceded by a visual stimulus, i.e., situations. These could serve as foci for minimizing communication barriers and as referents to the structural integrity of the instrument.

Based on the aforementioned statement of the problem, it is the purpose of this study to: 1) Develop an evaluation instrument for the discipline of biology relevant to secondary school instruction; 2) Describe the inherent qualities of the instrument and identify uses in psychological studies; 3) Classify selected operations by taxonomic category using a scheme analogous to that outlined in the Taxonomy of Educational Objectives: Cognitive Domain; 4) Administer the instrument to samples of students representing diversity of biology instruction and concomitantly assess performance at each represented taxonomic level; 5) Compare performance by taxonomic level as defined above, by background in biology, and by achievement level on a recognized criterion measure of critical thinking ability.

Methods and Procedures

Instrumentation

Based on the above rationale the Assessment of Cognitive Transfer--An Evaluation Instrument for Secondary Biology Teaching was developed. It was written using a branching program format centered around nine structurally related biological situations. Each frame required a verbal response by the student. The criteria reflected in the major frames of the program served as the basis for the forty-two item Assessment of Cognitive Transfer in Science Inventory. Some of the

behaviors measured included: observations relevant to stated hypotheses or research designs, generation of hypotheses given underlying assumptions, designing experiments, recalling concepts, predicting results, explaining phenomena and discovering relationships based on observations. To satisfy the items in the ACTS Inventory students had to elicit the minimum performance set for each criterion.

Each item in the ACTS Inventory was classified by cognitive category using a scheme unique to science but patterned after the Taxonomy of Educational Objectives. Handbook I: Cognitive Domain (2). The knowledge level was extracted from the "BSCS Grid for Test Analysis" by Klinkman (11). The higher levels designated as the processes of science represented segments of the scheme developed by Brown (4). The categories of cognition represented in the inventory were: knowledge, application, collection of data, analysis of data, withholds judgment, synthesis and evaluation as arranged in order of their assumed hierarchy.

The experimenter served as the sole arbiter in classifying each item in the ACTS Inventory by cognitive category. Table 1 contains the distribution of items by category of cognition for each of the 42 items in the inventory.

Sampling Procedure

Selection of students for the sample was based primarily on the BSCS Attitude Inventory by Blankenship (1) and the Biology Classroom

TABLE 1

DISTRIBUTION OF ACTS ITEMS

BY COGNITIVE CATEGORY

Category*	Item Number
1. Knowledge	6, 7, 12, 31
2. Application	1, 2, 3, 9, 11, 15, 16, 19, 20, 30, 34
3. Collection of Data	8, 14, 17, 22, 25, 32, 40
4. Analysis of Data	13, 21, 26, 29, 37, 39
5A. Synthesis of Data	10, 28, 33, 38, 41
5B. Withholds Judgment	4, 18, 23, 27, 35
6. Evaluation of Data	5, 24, 36, 42

*Descriptions of categories are given in Appendix D.

Activity Checklist by Kochendorfer (12). The population was limited to 1968-69 biology students of teachers enrolled in a BSCS Summer Institute at the University of Iowa during 1969. Students from four teachers were selected at random for interviewing. The students had BSCS or non-BSCS backgrounds as determined by teacher scores on BSCS Attitude Inventory and mean composite rating on BCAC. In addition one sample was selected from a group taking BSCS biology. The classes randomly sampled are characterized by the data in tables 2 and 3.

For the purpose of sampling only those teachers from Iowa schools were considered. Subsequently, all schools involved administer the Iowa Test of Educational Development (10). One subtest in this instrument served as a criterion measure of critical thinking for comparing performance on the ACTS Inventory. This subtest, Test 6: Ability to Interpret Reading Materials in the Natural Sciences, was described in the ITED manual for teachers and counselors:

> ". . Tests 5, 6, and 7 measure the ability to interpret reading materials in the social studies, the natural sciences, and literature. While constructed in the external form of a reading comprehension test, these three tests are designed to measure much more than generalized reading skills. Essentially, they are intended to measure the pupil's ability to do critical thinking in the broad areas designated. They are concerned not so much with what the pupil has learned, in the sense of specific information, but rather with how well he can use whatever he has learned in acquiring, interpreting and evaluating new ideas, in relating new ideas to old, and in applying broad concepts and generalizations to new situations or to the solution of problems . . ." (9)

TABLE 2

CLASSIFICATION OF SAMPLING GROUPS

	High Attitude Inventory SCAC	Low Attitude Inventory SCAC
BSCS	Class A	
Non-BSCS	Class B	Classes C, D
One Month BSCS	Class E*	

*No SCAC scores available.

TABLE 3

CHARACTERISTICS OF SAMPLE GROUPS BASED ON AVAILABLE DATA

Sample Designation	SCAC (Mean)	AI	WG	TOUS	Highest Degree	Year of Highest Degree	Years Teaching Experience
A	33.9	24	73	47	MA	1967	6
B	30.0	27	66	49	MA	1967	7
C	27.8	21	63	45	BA	1966	3
D	26.9	17	58	37	BS	1965	3
E	(Same teacher as Sample Group A, no student data currently available.)						

SCAC = Science Classroom Activities Checklist (12)

AI = Attitude Inventory (1)

WG = Watson-Glaser Critical Thinking Appraisal (19)

TOUS = Test On Understanding Science (5)

Statement of Null Hypotheses

Null Hypothesis for Independence of Classes With Respect to ITED Test 6.

1) The five classes do not differ with respect to the frequency of students in high, middle, and low levels of performance for ITED Test 6.

Null Hypotheses for Independence With Respect to Total ACTS Inventory Scores.

2) The five classes do not differ with respect to the frequency of students in the high, middle, and low levels of performance for the total ACTS Inventory.

3) Performance of students by level for ITED Test 6 does not differ with respect to performance of students by level for the total ACTS Inventory.

Null Hypotheses for Independence for Acts Inventory Scores by Cognitive Category.

4) The five classes do not differ with respect to performance by level for Categcry 1 (Knowledge) of the ACTS Inventory.

5) The five classes do not differ with respect to performance by level for Category 2 (Application) of the ACTS Inventory.

6) The five classes do not differ with respect to performance by level for Category 3 (Collection of Data) of the ACTS Inventory.

7) The five classes do not differ with respect to performance by level for Category 4 (Analysis of Data) of the ACTS Inventory.

8) The five classes do not differ with respect to performance by level for Category 5A (Synthesis of Data) of the ACTS Inventory.

9) The five classes do not differ with respect to performance by level for Category 5B (Withholds Judgment) of the ACTS Inventory.

10) The five classes do not differ with respect to performance by level for Category 6 (Evaluation of Data) of the ACTS Inventory.

11) The performance of students on the ACTS Inventory by levels does not differ when two contiguous categories are compared with other pairs of contiguous categories.

12) The performance of students on the ACTS Inventory by levels does not differ when one cognitive category is compared with the remaining cognitive categories.

Results and Interpretations

For empirical reasons reflected in the nature of this study and since the conditions for parametric tests could not be satisfied, non-parametric procedures were employed. The Chi-Square test for k independent samples (15, pp 175-179) was used to test most of the above null hypotheses. This test requires that the expected frequencies in each cell not be too small (15, p. 179). According to Tate (16, p. 71) how small is a difficult question to answer. There is general agreement that when the degrees of freedom is larger than two, fewer than twenty percent of the cells should contain an expected frequency of less than five. No cell should contain an expected frequency of less than one. However, according to Tate (50, p. 71) there is considerable evidence that this requirement is too high if: 1) there are two or more degrees of freedom or 2) the expected frequencies over the entire table average out to more than five per cell.

The intention here is not to emphasize the sheer significance of any test. Rather, in agreement with Hayes (8, p. 614), an attempt was made to appraise the strength of the relationships presented. Out of interest, design, and necessity all conclusions are based on the apparent predictive relationships in the data.

Tests for Independence of Classes

These tests determined if the classes of students that were randomly sampled were from the same or identical populations. Scores on ITED Test 6 were used to compare the classes. As indicated in Table 4, students were grouped by thirds on the basis of percentile ranks on Iowa Norms. Note that for all contingency tables the numbers in parentheses represent expected values whereas those without parentheses represent observed frequencies. The following χ^2 test for k independent samples was employed in determining independence (15, p. 175):

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad ; \quad df = (k-1)(r-1) \qquad (1)$$

Since $p < .30$ is greater than $\alpha = .05$, the null hypothesis was retained, i.e., the classes were independent with respect to scores on ITED Test 6. The diversity exhibited by the classes respective of the distribution of scores by levels was attributed to random sampling error.

Because of the limited power associated with the χ^2 test, it was decided that the significance of the diversity be assessed using the Kruskal-Wallis one-way analysis of variance by ranks. This test assumes that the variable being assessed has an underlying continuous distribution which can be measured at least ordinally. It should be stressed that the Kruskal-Wallis test has asymptotic efficiency of $3/\pi$ = 95.5 percent with respect to the F test as described by Andrews in Siegel (15, p. 194). Using formula (2) as given in Siegel (15, p. 185) the corrected H for the data in Table 4 was 8.273. Respective of the

$$H = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1);\ df = k-1 \quad (2)$$

df value associated with this statistic, the decision to retain the null hypothesis was verified although the value approached rejection at the 0.05 level. This value is actually closer to the 0.10 level of significance.

It appears from the contingency table in Table 4 that the classes do not reflect a normal distribution with respect to ITED Test 6. Furthermore, the frequency distribution of scores exhibited a skewed, bimodal form. The significance of the discrepancy from the normal distribution was determined by a χ^2 Goodness of Fit test (17, pp. 483-484). The hypothesis that the parent population was normally distributed was retained at the 0.05 level of significance (χ^2 = 14.69, df = 10).

However, the χ^2 test is somewhat insensitive to skewness and kurtosis because of its failure to regard the signs of the discrepancies.

TABLE 4

DISTRIBUTION OF STUDENT SCORES

ON ITED TEST 6 (IOWA NORMS) FOR EACH CLASS

ITED Level*	Class A	B	C	D	E	Total
High	13 (9.2)	11 (9.2)	8 (9.2)	7 (9.2)	7 (9.2)	46
Middle	6 (6.0)	3 (6.0)	5 (6.0)	8 (6.0)	8 (6.0)	30
Low	1 (4.8)	6 (4.8)	7 (4.8)	5 (4.8)	5 (4.8)	24
Total	20	20	20	20	20	100

*by thirds based on percentile rank

$\chi^2 = 10.63$
df = 8
$p < .30$
Decision: Cannot reject the null hypothesis

When the signs of the discrepancies appear to exhibit a pattern, the χ^2 test is not appropriate and the more sensitive alpha statistics should be used (17, p. 484). To test for normal distribution of the parent population, it is necessary to compute the α_3 and α_4 values (17, pp. 180-181) and compare the values with the tabled values for a normally distributed population (17, Table G). According to Tate (17, p. 447) the assumption of normal distribution is in doubt if either α value is significantly large. This assumption was reflected for these data. Although non-normal peakedness was not so sufficiently large as to discredit the hypothesis of normal distribution, this hypothesis was in doubt because of skewness to the left as evidenced by the α_3 value. The rationale developed in the above paragraphs precluded the use of parametric tests in the analysis of data (15, pp. 19-20).

Tests of Independence With Respect to Total ACTS Inventory Scores

Table 5 illustrates the test for independence between classes when compared on the basis of ACTS Inventory scores. The frequency distribution of scores on the ACTS Inventory for the combined classes was divided into thirds. Because of the disproportionate number of students at the extremes of some of the levels, students with borderline scores were randomly assigned to either of the adjacent levels to effect equal representation in the levels. The decision to retain the null hypothesis was evidenced by the indicated χ^2 value. The obvious discrepancy between the expected and observed scores for

TABLE 5

DISTRIBUTION OF STUDENT TOTAL SCORES

ON ACTS TEST FOR EACH CLASS

ACTS Level*	Class A	B	C	D	E	Total
High	11 (6.6)	7 (6.6)	6 (6.6)	4 (6.6)	5 (6.6)	33
Middle	5 (6.8)	5 (6.8)	4 (6.8)	11 (6.8)	9 (6.8)	34
Low	4 (6.6)	8 (6.6)	10 (6.6)	5 (6.6)	6 (6.6)	33
Total	20	20	20	20	20	100

*by thirds based on frequency distribution of ACTS scores

$\chi^2 = 13.33$
df = 8
$p < .20$
Decision: Cannot reject the null hypothesis

Class A students at the high ACTS Inventory level may be, in part, a reflection of the somewhat higher proportion of students in this class who were in the highest third on ITED Test 6. Since there were only 20 students in each class, the use of an ITED Test 6 control as a third dimension in a χ^2 test was negated. One would suspect on the basis of this outcome that instruction in biology as defined for each class does not alter achievement as measured by the ACTS Inventory.

By employing the Kruskal-Wallis test to the above, it was found that there is indeed no statistically significant difference at the 0.05 level in average scores on the ACTS Inventory for the five classes. The corrected H value was 7.447. The probability is less than 0.10 that this H value would be obtained in the case of a true null hypothesis.

As an assessment of concurrent validity Table 6, showing the relationship between ITED Test 6 scores and ACTS Inventory scores, was generated. The corresponding probability figure is less than 0.001, and the hypothesis of independence was strongly discredited. That is, if a student's score is in the high level for the ACTS Inventory it is highly likely that his score on ITED Test 6 will also be high. This same pattern is also in evidence for the low levels. In cases such as this where the null hypothesis was rejected a contingency coefficient was calculated as a measure of predictive association. This value was derived by using formula (3) as cited in Siegel (15, p. 197).

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \qquad (3)$$

The contingency coefficient C approximates r_{xy} as the number of categories for each variable increases. For a 3 X 3 table as used in Table 6 the computed C cannot exceed 0.816. Thus, the C value for Table 6 of 0.556 is respectably high. Since both tests (ACTS Inventory and ITED Test 6) measure aspects of critical thinking in science, a high C value is to be expected. It does support the concurrent validity of the ACTS Inventory. As evidenced by the extreme cells in Table 6, this coefficient could be assigned a plus sign.

Tests of Independence for ACTS Inventory Scores by Cognitive Category

A series of tests were run using formula (1) to determine independence of the classes respective to their distribution at each cognitive category. A summary of the probability figures associated with each of these tests is given in Table 7.

For Cognitive Category 1 (Knowledge) performance by level varied from class to class. In studying the contingency table it was revealed that the contingency coefficient was positive. Its value was calculated at C = .376. Several cells in this contingency table contributed disproportionately to the χ^2 value. The researcher was suspect of this outcome because Class E was at the time of the interviews enrolled in a biology course. Thus one would predict, as confirmed from the data, that this class would outperform the other

TABLE 6

DISTRIBUTION OF STUDENT TOTAL SCORES

ON ACTS TEST BY LEVEL ON ITED TEST 6

ACTS Level*	ITED Level# High	Middle	Low	Total
High	22 (9.9)	7 (11.2)	2 (9.9)	31
Middle	9 (9.9)	15 (11.2)	7 (9.9)	31
Low	1 (12.2)	14 (13.7)	23 (12.2)	38
Total	32	36	32	100

*by thirds based on frequency distribution of ACTS scores

#by thirds based on frequency distribution of standard scores on ITED

$\chi^2 = 44.74$
df = 4
$p < .001$
Decision: Reject the null hypothesis
C = .556

TABLE 7

ASSOCIATED PROBABILITY FIGURES

FOR EACH COGNITIVE CATEGORY WHEN

PERFORMANCE ON THE ACTS TEST

AMONG CLASSES WAS COMPARED

Cognitive Category	Probability Figure	Decision: α = .05
Knowledge	$p < .02$	Reject
Application	$p < .20$	Retain
Collection of Data	$p < .80$	Retain
Analysis of Data	$p < .30$	Retain
Synthesis of Data	$p < .50$	Retain
Withholds Judgment	$p < .20$	Retain
Evaluation of Data	$p < .10$	Retain

classes for the Knowledge category. The data further revealed that students from BSCS classes (Class A) or students with teachers using BSCS philosophy and rationale (Classes A and B) may out perform students with more traditional biology backgrounds. One could suggest that emphasis by a BSCS teacher on developing the structure of the discipline may tend to promote retention of facts, concepts and principles. Such a premise, however, would not guarantee that all students with BSCS training retain more knowledge. Possibly only the better students as defined by ITED Test 6 can profit in this respect. Furthermore, there was little difference in distribution for Classes C and D. This was to be expected since both classes were taught by teachers using similar philosophy and teaching strategies. Both were also non-BSCS classes (refer to Tables 2 and 3). Deletion of Class E from the contingency table resulted in failure to reject the null hypothesis. This borderline test indicates that further research is needed to confirm the results.

Note that for all the remaining categories represented in Table 7 the null hypothesis was retained. At first glance this suggests that biology background, independent of other measures of achievement or aptitude, does not alter performance significantly at the higher cognitive categories. This is especially true for the Collection of Data category for which the probability of obtaining such a χ^2 value under a true null hypothesis was about 0.80. At the other extreme the ability to evaluate data approaches rejection of the null hypothesis at the .05 level of significance. Here the χ^2 value is significant

at the 0.10 level. It may well be that a teacher using BSCS strategies could favorably influence a student's ability to evaluate. The range of p values given for these tests reflects trends that should encourage further research in this area. The data further suggest, because of the range in p values, that factors other than intelligence or native ability are operational.

The underlying null hypotheses reflected in Table 8 considered the independence of contiguous cognitive categories respective of other contiguous categories. The research hypothesis asked is whether students who do well on adjacent categories of cognition also do well on other combinations of adjacent categories.

Table 8 represents a three dimensional modification of the χ^2 test for independence. Formula (4) as referenced by Tate (16, pp. 74-75) was used for calculating the χ^2. Here expected frequencies are

$$\chi^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}} \quad ; \; df = abc - (a + b + c - 2) \quad (4)$$

calculated by finding the product of the three marginal totals for each cell and dividing by the square of the grand total.

In Table 8 the null hypothesis was strongly rejected since the corresponding probability figure was less than 0.001. This was further substantiated by the positive contingency coefficient of C = .534. Because rejection was indicated two dimensional tests that compared separate pairs of contiguous categories were performed. Rejection of the null hypotheses was also suggested in each case.

TABLE 8

DISTRIBUTION OF STUDENT SCORES ON COGNITION CATEGORIES 1 AND 2 RELATIVE TO SCORES ON COGNITION CATEGORIES 3 AND 4, AND COGNITION CATEGORIES 5 AND 6

Categories 1 and 2*	Categories 3 and 4*			
	High		Low	
	Cats. 5 and 6*		Cats. 5 and 6	
	High	Low	High	Low
High	23 (9.2)	5 (9.2)	9 (13.3)	8 (13.3)
Low	4 (11.3)	9 (11.3)	14 (16.2)	28 (16.2)

Totals: Categories 1 and 2 High, 45
Low, 55
Categories 3 and 4 High, 41
Low, 59
Categories 5 and 6 High, 50
Low, 50
Grand Total, 100

*levels determined from frequency distributions of the sums of the scores on each combination of categories

$\chi^2 = 39.97$
df = 4
$p < .001$
$C = .534$
Decision: Reject the null hypothesis

A series of tests were conducted to determine the extent to which performance on one cognitive category is predictive of performance on the remaining cognitive categories combined. The C values associated with these tests is given in Table 9. All null hypotheses were rejected. This suggests that performance at any one level is predictive of performance on the entire ACTS Inventory. One is then led to speculate that each category contains items varying in difficulty. Performance on Category 2 (Application) was the best predictor of success on the total inventory. Possibly this category is the least affected by student background and concomitantly most influenced by the general ability of the students. Unfortunately contingency tables could not be used to assess performance by class comparisons or by level comparisons on ITED Test 6 with any degree of confidence since the calculated expected frequencies in most instances did not fulfill the minimum requirements. In future research, where larger samples are employed, such comparisons should further elucidate the reasons for rejection of the null hypotheses found here.

Estimates of Item Difficulty, Discrimination and Test Reliability

Table 10 illustrates the item difficulty for the 42 item ACTS Inventory. Note that although not rectangular in distribution, a considerable spread in difficulty is evident. The values represented

TABLE 9

CONTINGENCY COEFFICIENTS FOR TESTS COMPARING SCORES ON EACH COGNITIVE CATEGORY OF THE ACTS TEST WITH SCORES ON THE REMAINING COGNITIVE CATEGORIES

CATEGORY COMPARISONS	CONTINGENCY COEFFICIENT VALUE*
1 VS 2-6	C = .422
2 VS 1,3-6	C = .551
3 VS 1,2,4-6	C = .377
4 VS 1-3,5,6	C = .482
5A VS 1-4,5B,6	C = .430
5B VS 1-4,5A,6	C = .354
6 VS 1-5	C = .420

*refer to Table 1 for category designations

**all values were positive based on the associated contingency table for each

TABLE 10

DIFFICULTY INDICES DISTRIBUTION

FOR THE TOTAL ACTS INVENTORY

Interval*	Number	Percent	Item Numbers
85 - 89	1	2	25
80 - 84	1	2	18
75 - 79	5	12	12, 21, 30, 38, 40
70 - 74	1	2	3
65 - 69	3	7	4, 20, 32
60 - 64	1	2	29
55 - 59	1	2	22
50 - 54	1	2	15
45 - 49	2	5	2, 11
40 - 44	3	7	17, 19, 27
35 - 39	2	5	6, 39
30 - 34	1	2	34
25 - 29	1	2	16
20 - 24	5	12	1, 14, 23, 33, 37
15 - 19	5	12	5, 8, 10, 31, 35
10 - 14	5	12	9, 24, 26, 28, 42
5 - 9	1	2	36
0 - 4	3	7	7, 13, 41

*Smaller index values represent greater difficulty

were determined by the number of students getting each item correct divided by the number of students. Thus, an index of difficulty is an expression of the number of correct responses for an item. Unfortunately, test construction has been dominated by theories and practices that stress the identification and measurement of individual differences (18). Items at the 50 percent level of difficulty are most effective in discrimination. Such a spread in difficulty as exhibited here has recently been recommended by Tyler (18) as a goal in developing better measuring instruments.

In Table 11 the item difficulty is expressed by cognitive category. As expected from the previous discussion the categories do exhibit a wide spectrum of difficulty with respect to the items in them. This trend, however, is not as apparent for the higher cognitive categories, i.e., 4, 5A and 6. Items in Category 5B (Reserves Judgment), required only a yes or no response. The guess factor present here could account for the diversity of difficulty shown. Respective of the other higher categories, especially Category 6 (Evaluation), it could be concluded that criteria so classified are less likely to be satisfied simply because, by definition, they represent the highest level of cognition.

The split-halves method was used to calculate indices of discrimination. They were determined by finding the difference in the proportion of correct responses between the groups of students scoring in the top 27 percent on the total ACTS Inventory and the bottom 27 percent (6, p. 352) Positive values indicate that high scoring

TABLE 11

DIFFICULTY INDICES DISTRIBUTION

FOR THE SEVEN COGNITIVE CATEGORIES

OF THE ACTS INVENTORY

Index Interval	Cognitive Category 1	2	3	4	5A	5B	6
85 - 89			25*				
80 - 84						18	
75 - 79	12	30	40	21	38		
70 - 74		3					
65 - 69		20	32			4	
60 - 64				29			
55 - 59			22				
50 - 54		15					
45 - 49		2,11					
40 - 44		19	17			27	
35 - 39	6			39			
30 - 34		34					
25 - 29		16					
20 - 24		1	14	37	33	23	
15 - 19	31		8		10	35	5
10 - 14		9		26	28		24,42
5 - 9							36
0 - 4	7			13	41		

*item numbers are recorded in the table

students answered the item correctly more frequently than low scoring students. The discrimination value approaches one as the relative difference increases. Table 12 shows the discrimination for each item on the total ACTS Inventory and for each cognitive category respectively. Interestingly, no negatively discriminating items are evidenced. In other words, there were no items on which students with low scores outperformed students with high scores.

As a final note, the reliability of the ACTS Inventory as computed by a modification of Flanagan R as cited in Ebel (6) was 0.86. A minimum value of 0.70 is recommended for such measures of reliability.

Conclusions

An instrument was developed that measured cognitive performance in biology at seven levels of cognition. Based on results obtained from the ACTS Inventory, the criterion measure extracted from the instrument, and ITED Test 6, a measure of critical thinking ability in science, the following conclusions were formed. They reflect the validity of the instrument and suggest directions for future research.

1) The five classes tested did not differ with respect to performance on the total ACTS Inventory, i.e., the ratios of students performing at the high, middle, and low levels for the ACTS Inventory did not vary significantly from class to class. Thus, the null hypothesis could not be rejected. When total ACTS Inventory scores served as a criterion measure, students taught by BSCS teachers or by teachers

TABLE 12

DISCRIMINATION INDICES DISTRIBUTION

FOR THE SEVEN COGNITIVE CATEGORIES

OF THE ACTS INVENTORY

Index Interval	Cognitive Category 1	2	3	4	5A	5B	6
.70 - .79			14*	29	10		
.60 - .69	6,31	19	22	21	33	4,18	
.50 - .59		1,11 15,20 34	32,40	37,39	28,38		24,42
.40 - .49	12		17	13		23,27	5,36
.30 - .39		2,16 30					
.20 - .29	7		8,25	26	41		
.10 - .19		3,9				35	

*item numbers are recorded in the table

using BSCS philosophy and rationale do not perform significantly better than students taught by non-BSCS teachers who do not use BSCS philosophy and rationale. It should be noted that in most cases the ACTS Inventory was administered at least three months after the students had completed their biology course.

2) Performance of students by level on ITED Test 6 was associated with performance of students by level on the total ACTS Inventory. Thus, the null hypothesis was rejected.

3) With one exception when levels of performance on the individual cognitive levels were used to compare the five classes, the null hypotheses could not be rejected. The null hypothesis was rejected for Category 1 (Knowledge). However, when the class presently taking biology was deleted, the null hypothesis could not be rejected at the 0.05 level of significance although the classes could be considered dependent at the 0.10 level of significance. This same level of significance was found for the χ^2 value associated with the table comparing performance by classes on Category 6 (Evaluation). The greatest independence among the five classes was shown for Category 3 (Collection of Data).

4) Performance by levels for pairs of contiguous categories of cognition for the ACTS Inventory in the assumed hierarchy of cognition was related to performance on other contiguous pairs of cognitive categories. Thus, the null hypotheses were rejected. Students who performed at one level on adjacent categories of cognition tended to perform at the same level on other combinations of adjacent categories.

Although diversity in the range of association existed, it was apparent that performance by students was consistant throughout contiguous pairs of categories.

5) Performance by level at any one cognitive category of the ACTS Inventory is related to performance by level on the remaining cognitive categories combined. That is, the underlying null hypotheses of independence were rejected. Performance on Category 2 (Application) was the best predictor of success on the total ACTS Inventory.

6) The test analysis data for the ACTS Inventory corroborated the above findings that performance by levels for one or a combination of ACTS Categories is independent of performance on other categories. This could in part be attributed to the spread of difficulty exhibited for most of the cognitive categories, i.e., items in most categories ranged from easy to difficult. This trend was less evident for the higher cognitive categories. No item on the ACTS Inventory had a negative index of discrimination. The reliability for the ACTS Inventory, determined by a modification of the Kuder-Richardson formula was 0.86.

In the light of the above findings it is evident that the scheme developed here for evaluation in biology has a multiplicity of applications. A propensity toward administration of the data collecting instrument via computer assisted instruction is indeed evident. The technology in this area has already been developed. Large scale testing could well provide the means for assessing the purported goals of science education.

Literature Cited

1. Blankenship, J. W., "The Development of an Attitude Inventory Designed to Determine Reactions of Biology Teachers to BSCS Biology", Research and Curriculum Development in Science Education 1, A. E. Lee, Ed., Science Education Center, The University of Texas, Austin, 1968, pp. 21-28.

2. Bloom, B. S., et. al, (Editors), Taxonomy of Educational Objectives--The Classification of Educational Goals--Handbook I: Cognitive Domain, David McKay Company, Inc., New York, 1956.

3. Bloom, B. S., "Testing Cognitive Ability and Achievement", in Handbook of Research on Teaching, N. L. Gage, Ed., Rand McNally and Co., Chicago, 1963, pp. 379-397.

4. Brown, W. R., "Defining the Processes: A Statement by Educators in India", The Science Teacher, Vol. 35, 26 (December, 1968).

5. Cooley, L. E., and L. E. Klopfer, Test On Understanding Science, Form W., Educational Testing Service, Princeton, N. J., 1961.

6. Ebel, R. L., "Procedures for the Analysis of Classroom Tests", Educational Psychol. Measmt., Vol. 14, 277 (1954).

7. Ennis, R. H., "Needed Research in Critical Thinking", Educ. Leadership, Vol. 21, 17 (1963).

8. Hayes, W. L., Statistics For Psychologists, Holt, Rinehart and Winston, New York, 1963.

9. How to Use the Test Results of the Iowa Tests of Educational Development. Tenth Ed., The University of Iowa, Iowa City, 1969.

10. Iowa Tests of Educational Development, Forms Y4 and X4, Scientific Research Association, Chicago, 1960.

11. Klinckman, E., "The BSCS Grid for Test Analysis", Biological Sciences Curriculum Study Newsletter, Sept. (1964), pp. 17-21.

12. Kochendorfer, L. H., "The Development of a Student Checklist to Determine Classroom Teaching Practices in High School Biology", in Research and Curriculum Development in Science Education 1, A. B. Lee, Ed., Science Education Center, The University of Texas, Austin, 1968, pp. 71-78.

13. Lisonbee, L., "Testing, What For?", The Science Teacher, Vol. 33, 27 (May, 1966).

14. Ramsey, G. A., and R. W. Howe, "An Analysis of Research on Instructional Procedures in Secondary School Science, Part I--Outcomes of Instruction", The Science Teacher, Vol. 36, 62 (March, 1969).

15. Siegel, S., Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill Book Co., New York, 1956.

16. Tate, M. W. and R. C. Clelland, Nonparametric and Shortcut Statistics, Interstate Printers and Publishers, Inc., Danville, 1957.

17. Tate, M. W., Statistics in Education, The Macmillan Co., New York, 1955.

18. Tyler, R. W., "Resources, Models, and Theory in the Improvement of Research in Science Education", J. of Research in Science Teaching, Vol. 5, 43 (1967).

19. Watson-Glaser Critical Thinking Appraisal, Revised Form Zm, Harcourt, Brace and World, Inc., New York, 1961.

APPENDIX

ASSESSMENT OF COGNITIVE TRANSFER

IN SCIENCE INVENTORY

(With Explanatory Notes)

1. Elicits through direct and indirect questioning three components of the petri dish related to the growth of living organisms. (1-A2, 1-A3, or 1-A4)*

   After examining the petri dish the student should list a minimum of three components of the solid material in the bottom of the dish. The following are deemed appropriate: agar, gelatin-like substance, food, protein, minerals, starch, sugar, vitamins, water, nutrients, carbohydrates, fats, lipids, organic matter, etc.

2. States that the dishes allow for a free exchange of gases. (1-B1, 1-B2, or 1-B3)

3. States that the dishes provide an environment for growing pure cultures. (Cover prevents entrance of other organisms.) (1-C1)

   The student's statement should reflect the fact that he is aware that when such dishes are left with the cover off spores from the air can enter and grow.

4. Reserves judgement when asked if the same organism is growing in each dish. (2B1)

   To satisfy this criterion the student must indicate that he is not sure that the same thing is growing on each plate. This is of course dependent upon his conception of "same thing". All such decisions in the ACTS Inventory are to be classified as Reserve Judgement. The evaluator need not concern himself with these criteria.

5. Gives reason for reserving judgement. (2-B2a, 2-B2b, or 2-B3)

   To satisfy this criterion the student must state that he does not have adequate information to make a decision. By stating that he is not sure because the organisms on the two plates look different does not satisfy the criterion.

*Refers to frame designations in interview sequence.

6. Given examples thereof recalls the term species. (2-C1)

7. Defines species in terms of reproductive isolation. (2-C2 or 2-C7)

   The student's definition must include a component that suggests that his concept of species encompasses reproduction isolation, i. e., he might state that to be of the same species organisms must be able to reproduce (reproduce fertile offspring). He does not satisfy this criterion by stating that species look similar or that they live in the same surroundings.

8. Suggests a procedure to determine if the plates contain the same species. (2-C3, 2-C5, or 2-C6)

   At this point the student has elicited or has just been given a definition of species. To satisfy this criterion the student should state what one could mate or cross the materials in each dish.

9. Predicts what results would be needed to confirm hypothesis. (2-C4)

   The student has either elicited or has just been told that one approach would be to mate or cross the materials in the two dishes. The student should respond by stating that he would look for fruits or offspring.

10. Relates the presence of fruits in Situation III to the experimental design of Situation II by stating that the presence of fruits in the experiment would confirm the hypothesis. (3-A1)

    To satisfy this criterion the student should state the structures he sees here might be the offspring produced by the mating of the two organisms in Situation II.

11. When given additional information satisfies Criterion 10 above. (Automatically satisfies this criterion if Criterion 10 is satisfied.) (3-A2b)

    All students are now given frame 3-A2. If the student did not satisfy Criterion 10, he is not asked frame 3-A2b. To satisfy this criterion requires the same response as Criterion 10 required.

12. When given an example thereof states a definition of the term <u>hypothesis</u>. (3-A4)

A satisfactory response would include a phrase that suggests subsequent verification by experimentation. A sample definition of hypothesis is given in Frame 3-A6.

13. Refines previous experiment by employing the stated definition of species in order to confirm that the two plates contain the same species. (3-A5)

    A statement analogous to that given in Frame 3-A8 would satisfy this criterion.

14. Given additional information satisfies Criterion 13 above. (Automatically satisfies this criterion if Criterion 13 is satisfied.) (3-A7)

15. Interprets results of experiment by relating to preformed hypothesis. (3-B2)

    Here the student's response must include a statement regarding the fertility of the initial offspring.

16. Explains the failure to produce fruits when two spores from two separate fruits are mated. (3-C2)

17. States (3) three morphological differences between plates of Situations III and IV. (4-A2 or 4-A5)

18. Reserves Judgment. (4-B1)

19. Provides evidence to support given assumption. (4-B2 or 4-B4)

20. Asks (3) relevant questions regarding enviornmental factors in which the plates were grown. (4-C1, 4-C5, or 4-C6)

    See Frame 4-C6 for example. Other questions may include such factors as: light, temperature, position of plates, humidity, added chemicals or atmosphere in plates.

21. Focuses on new problem by designing an experiment to determine the effect of light on fruiting. (5-A1)

    Here the student is asked to focus on a new problem. To satisfy this criterion the student must state that he would grow some plates in the light and some in the dark, In addition he must include any one of the following components or any other relevant component.

    1. continuum of light intensity, quality or direction.

2. genetic continuity of organisms used.
3. two or more variables simultaneously, one of which is light.
4. all variables constant except light.

22. Observes that only those plates grown in the light produce fruits. (6-A1)

The above statement should be included as part of the student's initial remarks. Any alternative response must be directly related to the effects of light on fruiting, i. e., his observations must be relevant to the stated design of the experiment.

23. Reserves Judgment. (6-A3 or 6-A5)

24. Gives reason for reserving judgment. (Dependent upon satisfying Criterion 23) (6-A6)

This criterion may be satisfied by proposing a specific mechanism for fruit production in the dark. That is, light may activate the synthesis of a specific enzyme necessary for fruit formation. Possible the synthesis of this enzyme could be controlled chemically.

25. Observes difference in vegetative growth between plates grown in the light and in the dark. (6-B1)

Here the students are asked to refocus their attention. The only apparent difference between the two plates is in the relative sizes of the fungal mats. They could state that the culture grown in the dark is larger or more dense than the one grown in the light.

26. When shown experimental results, induces relationship between two variables, i.e., light and growth rate and how they are related to the fruiting process. (6-B2)

Note that the student is asked to derive a relationship involving three factors. He does not satisfy this criterion by merely stating that fruits are produced in the light and not in the dark and growth is more rapid in the dark than in the light.

27. Reserves Judgment. (6-B3)

28. Develops alternative explanation for the production of fruits. (6-B4)

Other than the sample response the student could state that

possibly more time is required for fruiting to occur in the dark or that light triggers a sequence of chemical reactions that could also be accomplished by adding the right chemicals.

29. Given that an inhibitor to fruiting is normally produced by this organism, explains the relationship between the inhibitor and light. (6-B5)

30. Employs the principle of parsimony in deciding on the direction of further experimentation. (6-B7)

31. Suggests what the inhibitor might be when given that it is a normal product of respiration. (6B8, 6-B9, or 6-B11)

    Given that the product might be a simple product of respiration, the student predicts what this product might be. Satisfying this criterion would be premised on a fundamental understanding of the process of respiration.

32. In observing the results of this experiment states that the sealed plates in the light do not contain fruits. (7-A1)

    Although other observations are possible, this is the only one relevant to the stated experimental design.

33. Proposes relevant explanation for why sealed plates in the light did not produce fruits. (7-A2)

34. Given added information satisfies Criterion 33 above. (Automatically satisfies this criterion if Criterion 33 is satisfied.) (7-A5)

35. Reserves Judgment. (7A3)

36. Gives reason for reserving judgment that involves removal of inhibitor. (Dependent upon satisfying Criterion 35) (7-A4a)

37. Suggests procedure for determining if CO_2 is the substance that inhibits fruiting. (7-B1 or 7-B4)

    No added information is given in Frame 7-B4. The problem is merely restated for the student.

38. Predicts that if a substance could be placed in the sealed plates to absorb the CO_2, fruits would be produced if CO_2 was the inhibitor. (Automatically satisfies this criterion if Criterion 37 satisfied) (7-B5)

39. Given that KOH in solution is an effective CO_2 absorber, designs an experiment to determine what factors influence fruiting. (7-B2 or 7-B6)

    To satisfy this criterion the student must design an experiment using KOH in sealed plates in the light and in the dark.

40. In rethinking previously stated relationship between light and fruiting states that light is not necessary for fruiting to occur. (8-A1)

    This statement corresponds to the observation that sealed plates containing KOH grown in the dark contain fruits.

41. Provides relevant explanation (mechanism) for the appearance of fruits by interrelating all data collected to this point. (8-A2)

    Here the student must elicit relationships involving all the components of the criterion. (See sample response in interview sequence)

42. Interprets new situation involving different species of organism. (9-A1)

    To satisfy this criterion the student must state that possibly two strains of the same species were grown on this plate. They grew toward one another and mated. The line of fruits was then produced at the point of contact between the compatible strains. In satisfying this criterion the student must use several concepts developed throughout the context of the interview sequence. At the minimum performance level the student would state that two strains were mated and produced offspring.